Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 1181–1188

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Named Entity Recognition using Conditional Random Fields

"Nita Patil, Ajay Patil and, B. V. Pawar" *

"School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharshtra University, Jalgaon-425001(MS), India"

Abstract

Identifying named entities (NEs) present in electronic newspapers in regional languages is an important step in machine translation and summarization systems. In this paper, we propose a statistical named entity recognition system based on machine learning for the identification and classification of named entities present in Marathi language text. In our system, named entities are identified and classified using conditional random fields (CRFs). As being a morphologically rich language, statistical algorithms achieves good NE identification and classification accuracy but needs extra knowledge to improve accuracy. Experiments conducted on the FIRE-2010 corpus show that our system submitted for the challenge achieves the precision, recall and F1-measure of 82.33%, 70.68% and 75.51% under the CRF algorithm.

Keywords: Named Entities, NER, Machine Learning, Mallet, CRF, Conditional Random Fields

1. Introduction

More and more regional language text is generated with the development of electronic media. However, the information present in the text cannot achieve expected outcome if directly used by natural language processing tasks. Highlighted named entities can add structure to the unstructured raw text, which can be freely used by focused

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: nvpatil@nmu.ac.in

1877-0509
10.1016/j.procs.2020.03.431

information retrieval, machine translation and summarization systems. Named entity recognition (NER) tasks includes identification and classification of word or phrase. The example below illustrates the NER task:

पंडित मास्टर दीनानाथ मंगेशकर एक शास्त्रीय गायक आणि थिएटर भूमिकेत होते. 2001 मध्ये, लताला 'भारतरत्न' भारताचा सर्वोच्च नागरी पुरस्कार प्रदान करण्यात आला. त्या वर्षी, लता मंगेशकरांनी, पुणे येथे मास्टर दीनानाथ मंगेशकर हॉस्पिटलची स्थापना केली. हॉस्पिटलचे व्यवस्थापन लता मंगेशकर मेडिकल फाऊंडेशन द्वारे करण्यात येते.
Pandit Deenanath Mangeshkar was a classical singer and theatre actor. In 2001, Lata was awarded the Bharat Ratna, India's highest civilian award. That year, Lata Mangeshkar established the Master Deenanath Mangeshkar Hospital in Pune. The management of the hospital is done through the Lata Mangeshkar Medical Foundation.

Processing of above sentence detects nine different NE instances i.e. पंडित मास्टर दीनानाथ मंगेशकर (Multi Word Person), 2001 (Year), लताला (Inflected Person), भारतरत्न (Miscellaneous Name), भारताचा (Inflected Location), लता मंगेशकरांनी (Multi Word Inflected Person), पुणे (Location), मास्टर दीनानाथ मंगेशकर हॉस्पिटलची (Inflected Multi Word Organization), लता मंगेशकर मेडिकल फाऊंडेशन (Multi Word Organization)

During the last decade, a large amount of research has been carried out to face a challenge and many kinds of systems have been developed for NE recognition. Named entity recognition is the significant information extraction task that must be developed for all languages in the world and for almost all domains. However, the task differs by the language, domain and approaches of system development. Some research in this field have focused on system development for one language and its evaluation on other languages to show the language and domain independency. Patil et al. [1] discussed the challenges in development of named entity recognition system for Marathi. Authors have stated that NER development is difficult for Marathi as it is morphologically rich and free word order language. In this paper, twelve main categories with forty named entity tags are defined. In this paper, we described the system based on machine learning algorithm conditional random fields (CRF). The paper is divided into four sections. The first section focuses on introduction and the need of the research. The second section reviews the research done for named entity recognition using CRFs. The third section describes the architecture of the NER system using CRFs. Final section describes the performance of the system developed for Marathi NE recognition.

2. Related Work

Peng et al. [2] has implemented a system that extracts information present in headers and references in research papers using CRF. Authors have investigated issues of regularization using Gaussian theory and focused on efficient use of feature space with CRF. The overall accuracy reported was 93.1% (HMM), 98.3% (CRF), 92.9% (SVM) for header information extraction and 85.1% (HMM), 95.37%(CRF) for extraction of paper references. Authors claim to reprove that CRF performs better than HMM. Shishtla et al. [3] implemented a NER system using CRF on the data released by NERSSEAL 2008. The corpus consists of training and testing datasets where the training dataset consists of 46,068 tokens (having 8,485 NEs), and the test data consists of 17,951 tokens (having 2,407 NEs). The BIE annotation scheme is used for tagging of the corpus. Language independent features such as previous and next word information, affix information; and language-dependent features such as part of speech tags, word chunks are used to identify NEs. The F-measure of 89.8% is reported for the best performing model. Luo et al. [4] have implemented joint model for entity recognition and disambiguation (JERL) that extracts NEs in text and links them to a knowledge base using CRF technique. Authors have evaluated JERL with CoNLL03 dataset and reported 91.2% accuracy in extracting NEs and linking them to knowledge base. Lample et al. [5] presented neural architectures for NER using small amount of supervised training data and unlabelled corpora. Authors have implemented two models for NER. First model is a combination of bidirectional long short-term memory networks with a CRF (LSTM-CRF) and second model is using transition-based parsing algorithm with states (S-LSTM). Authors have compared performance reported by both models with other existing systems for CoNLL03 dataset for two languages (English and German) and CoNLL02 dataset for Dutch and Spanish languages. Among both models, LSTM-CRF performed better with accuracy of 90.94% (English), 78.76% (German), 81.74% (Dutch), and 85.75% (Spanish). Srikanth and Narayana have developed CRF based noun tagger, trained on manually tagged dataset of 13,425 words and tested on dataset of size

6,223 words. The system based on CRF has reported 92% F-Score [6]. Li Wei et al. [7] have developed named entity recognition using conditional random fields and feature induction for Hindi language. Ekbal et al. [8] [9] has reported CRF based named entity recognition system for Bengali and Hindi language. Vijayakrishna et al. [10] has reported domain focused named entity recognition system for Tamil language using conditional random fields. We have proposed a basic NER system (Mner-CRF) for Marathi language using manually annotated NE tagged corpus comprising of 27,177 sentences.

3. System Architecture

CRF is relational learning model. NER using CRF is based on undirected graphical model of conditionally trained probabilistic finite state automata. CRF is used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. It incorporates dependent features and context dependent learning. It allows representing dependencies on previous classifications in a discourse. The basic idea is context surrounding name becomes good evidence in tagging another occurrence of the same name in a different ambiguous context. Most of the research in NER using CRF implementation needs feature vector consists of language features and POS tags, morphological analyzer, gazetteers, and NE annotated corpus.

For machine learning, the data needs to be converted to a feature vector for every word where the feature map contains context information of the word. The system (Mner-CRF) implemented feature function that takes a sentence, position of the current word, label of the current word and label of the previous word as a input parameters to the function from the training dataset and predicts most appropriate NE tag for a word in the sentence based on the learning. Training of CRF system is done with training dataset which returns a model file as an output. The model file is used by the CRF system in the testing process. Training CRF and generating model file is a core module in NER system development. Conditional random fields algorithm uses sentence x as the input sequence , where $x_{1,}x_2 , \dots . , x_m$, are the words of a sentence and tag sequence t , where $t_{1,}t_2 , \dots . , t_{m,}$is the sequence of output states, i.e. the NE tags. In

CRFs, the conditional probability $P(t_{1,}t_2 , \dots . , t_{m,} |\ x_{1,}x_2 , \dots . , x_m)$ is modeled by defining a feature map that maps an entire input sequence x paired with an entire state sequence t to some d-dimensional feature vector. Then we can model the probability as a log-linear model with the parameter vector. To apply this model, the data having 27,177 Marathi sentences containing 63,236 unique words is loaded. We have used the conditional random field (CRF) implementation provided by Mallet [11]. Mallet is a statistical package for statistical natural language processing. Mallet includes tools for named entity extraction from text. Mallet provides implementation of linear chain CRF that can be used for sequence tagging. CRF implementation includes three parts, model, trainer, and evaluator. Probability of output tag sequence for given input word sequence is computed. Marginal probability over states for given input sequence is computed using forward and backward algorithm. Figure 1 describes the whole NER process.

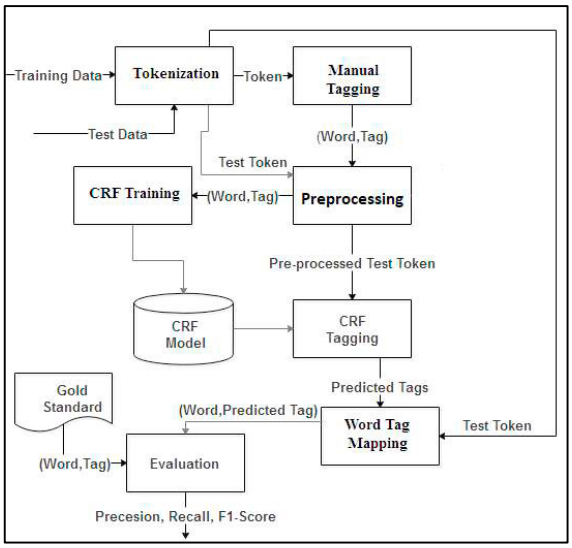

Fig.1. The NER System (Mner-CRF) Architecture

Training data is combined collection of 2000 news stories taken from FIRE-2010 [12] corpus. The collection of Marathi sentences is manually tagged. The preprocessed NE tagged data is given to CFR module for training, which generates CRF model file after exhaustive learning process. The CRF model file used for NE tagging of text, which is evaluated further for precision, recall and F1-Measure using gold standard dataset. The test dataset is given to splitter module that separates tags and words forms. These separated words forms are used as test dataset. The test dataset is given to CRF to compute output tag sequence for given input sequence in the test dataset based on the training given. CRF module generates output tag sequence which is further merged with words forms present in test dataset by merger module. Finally, merged output sequence which consists of words forms from test dataset and tags generated by CRF module is compared against held out data set for performance evaluation of the system.

3.1. Dataset Preparation

Twelve different types of NE classes are considered in this study viz. person, location, organization, miscellaneous, amount, number, measure, date, time, weekday, month, and year. Among these NE categories weekday, month and year are single word NE types and remaining are multiword NE types. Marathi news text corpus (FIRE 2010) is manually annotated for 40 different tags using IOBES encoding scheme. The corpus contains 27,177 Marathi sentences with an average number of 15 words in each sentence. The FIRE corpus constructed using news documents is used for both training and testing the system. The news documents are tokenized by a tokenizer resulting into word forms that are merged into a single file while preserving all punctuations and sentence separators. All word forms are tagged using 40 different tags manually. The dataset is partitioned into two non-overlapping parts in 80:20 proportion. First part (80%) is used to train the CRF models and second part (20%) is kept as held out data set. The number of NE instances present in testing corpus is shown in figure 2.

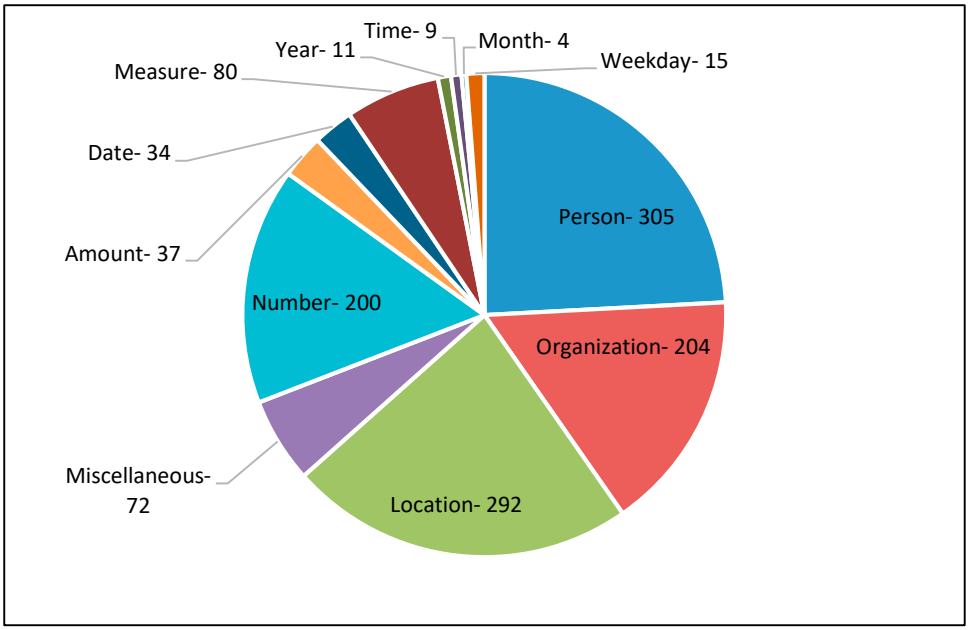

Fig. 2. Test Data Details

4. Results and Discussion

Little work has been reported in NER for Indian languages particularly for Marathi. Annotated corpora are not yet available in Marathi in required quantity and quality. In this research we therefore start with our experiments in building a CRF (Conditional Random Fields) based basic NE Tagger trained on a manually tagged data of 26,462 sentences and tested on a test data set of 715 sentences. The experiment results for held out data validation test are summarized in Figure 3.

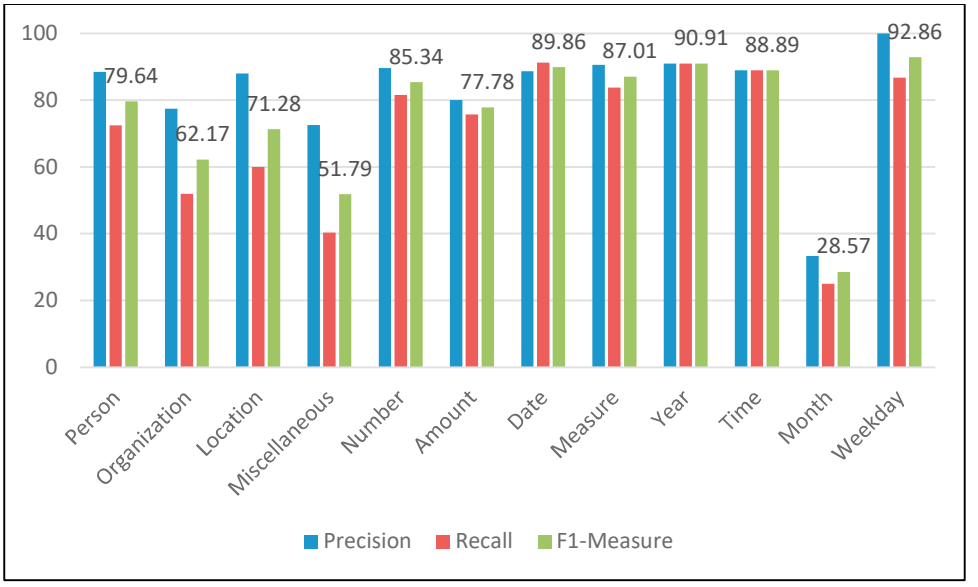

Fig. 3. Held out data validation test

The training dataset contains reasonable amount of person, location and organization tags. The probabilistic model of CRF in NER system performed well in recognition of these entities. It is found while annotating that many instances of these NEs were present in inflected forms. This has affected the recognition rate. Miscellaneous entity is too much diverse so that it is very much difficult to predict by any statistical NER system. Inflections were also observed in number, amount, measure and time NEs; still the Mner-CRF system has reported satisfactory recognition. Date, year and weekday NEs were recognized more precisely. It needs to be investigated the less recognition of month NE. Figure 4 shows precision recall and F1-Measure computed with and overall NE identification and average NE classification accuracy. The performance of proposed Mner-CRF system is shown in figure 4 reported 75. 7% (F1-Measure) overall named entity identification and 75.51% (F1-Measure) average named entity classification accuracy.

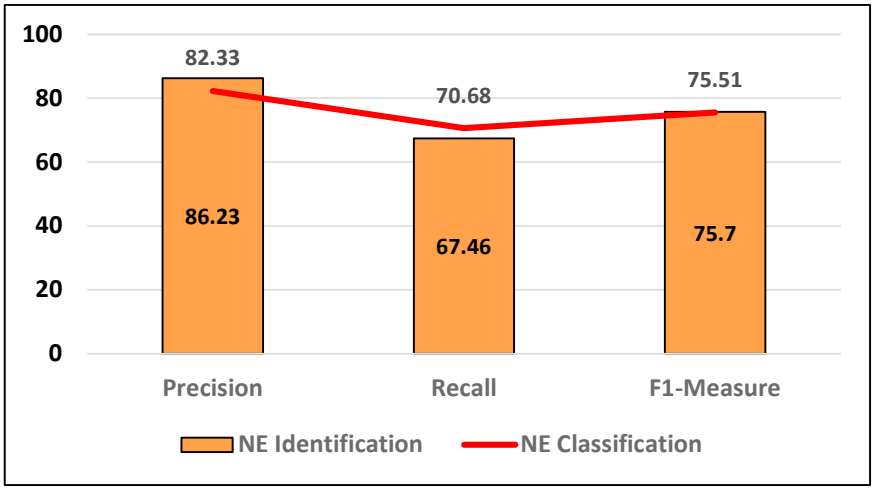

Fig. 4. overall NE identification and average NE classification accuracy

System performance is computed using held out data validation test are highly dependent on the training and test data division. System performance may affect if data that is significant for training, but it is part of testing. Therefore, 10-fold cross-validation test is also performed to evaluate the performance of the Marathi named entity recognizer. The annotated data is partitioned in 10 equal parts. Ten test runs have performed by keeping one part as test data and remaining 9 parts as training dataset. Figure 5 shows precision recall and F1-Measure computed with 10-fold cross-validation test. A number of experiments have been carried out in order to find out the most suitable features for NER in Marathi.

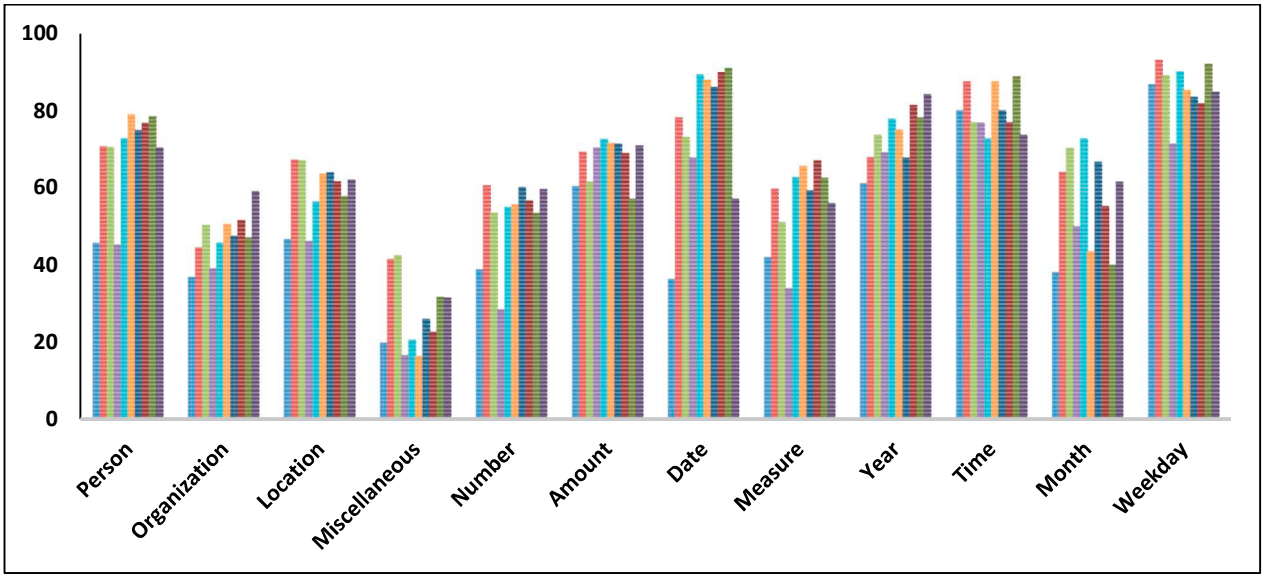

Fig. 5. 10-fold Cross Validation Test

Existing research reported in named entity recognition for Marathi language using CRF till date is shown in table 1.

Table 1. Existing NER systems for Marathi NER using CRF.

Author(s)	Domain Language	Training Datatype	Features	F1-Measure
Patawar et al.	English, **Marathi**	1000 tweets	Language independent	50.8%
Anup et. al.	Hindi, **Marathi**	3,884 sentences	WARMR rules	61.60%
Proposed (Mner-CRF)	**Marathi**	27,177 sentences	Language dependent	75.51%

Patawar et al. [13] reported a CRF based NER system for two languages (English and Marathi) tweets. Language independent features were used in building the NER system. 1000 political tweets were considered, out of which 800 tweets were used for training and 200 tweets were used for testing the system. 50.8% F1-Measure is reported for named entity recognition in Marathi tweets. Authors have stated that the performance of the system for Marathi tweets can be increased by incorporating with gazetteers, but gazetteers for Marathi names are not available. Anup et al. [14] developed hand-crafted rule based NER system for two languages (Hindi and Marathi). 3,884 sentences in Marathi were used in four different approaches implemented for Marathi NER. Authors have reported 61.60% F1-Measure (best among four approaches) for WARMR CRF based NER system.

5. Conclusion and Future work

In this paper the named entity recognition system for Marathi language based on machine learning algorithm conditional random fields is reported. Supervised machine learning based approach requires NE annotated data for training the system. The dataset comprising of 27,177 manually NE tagged sentences used to train and evaluate the system. Performance reported by NER system implemented is satisfactory. The challenge of improving recognition rate is introduced due to the rich morphology and inflected nature of Marathi language. The techniques reported in literature such as stemming, lemmatization etc. can be studied. It is needs to be researched that which technique can be easily implemented and more suited to handle the inflections in Marathi text to enrich the research in Marathi named entity recognition. The Mner-CRF system is also to be evaluated for unknown NEs present in test dataset.

References

[1] Nita Patil, Ajay S. Patil, and B. V. Pawar (2016) "Issues and Challenges in Marathi Named Entity Recognition" International Journal of Natural Language Computing (IJNLC), 5(6):15-31.

[2] Fuchun Peng, Andrew McCallum (2006) "Accurate Information Extraction from Research Papers using Conditional Random Fields" Information Processing and Management: an International Journal 42(4): 963-979.

[3] Praneeth Shishtla, Karthik Gali, Prasad Pingali, and Vasudeva Varma (2008) "Experiments in Telugu NER: A Conditional Random Field Approach" In Proceedings of the Workshop on NER for South and South East Asian languages (IJCNLP-08), Hyderabad, India:105–110.

[4] Luo Gang, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie (2015) "Joint named entity recognition and disambiguation" In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal: 879–888.

[5] Lample Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (2016) "Neural architectures for named entity recognition", In Proceedings of NAACL-2016, arXiv preprint arXiv:1603.01360 (2016).

[6] P. Srikanth and Kavi Narayana Murthy(2008) "Named Entity Recognition for Telugu", in Proceedings of the Workshop on Named Entity Recognition for South and South East Asian Languages, Third International Joint Conference on Natural Language Processing (IJCNLP-08), Hyderabad, India: 41-50.

[7] Li Wei, and Andrew McCallum (2003) "Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Induction", ACM Transactions on Asian Language Information Processing (TALIP) 2(3): 290-294.

[8] Ekbal, Asif and Bandyopadhyay, Sivaji (2009) "A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi", Linguistic Issues in Language Technology 2(1): 1-44.

[9] Ekbal, Asif, Rejwanul Haque, and Sivaji Bandyopadhyay (2008) "Named Entity Recognition in Bengali: A Conditional Random Field Approach", IJCNLP-2008.

[10] Vijayakrishna R, and Sobha L. (2008) "Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields", in Proceedings of the IJCNLP-08 workshop on NER for South and South East Asian Languages, Hyderabad, India. pp. 93–100.

[11] [http://mallet.cs.umass.edu date: 10 October 2016]

[12] [http://www.isical.ac.in/~fire/2010/]

[13] Maithilee L. Patawar, M. A. Potey (2016) "Named Entity Recognition from Indian tweets using Conditional Random Fields based Approach", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 5(5): 1541-1545.

[14] Anup Patel, Ganesh Ramakrishnan and Pushpak Bhattacharya (2009) "Incorporating Linguistic Expertise using ILP for Named Entity Recognition in Data Hungry Indian Languages", Proceedings of the 19th International Conference on Inductive Logic Programming (ILP'09), Leuven, Belgium: 178–185.